動詞型連体修飾表現の " N_1 の N_2 " への言い換え

片岡 明 増山 繁 山本 和英 † †

動詞を含む連体修飾表現を " N_1 の N_2 " という表現に言い換える手法を提案する. 動詞を含む連体修飾節は, 各文を短縮する既存の要約手法において, 削除対象とされている. ところが, 連体修飾節の削除によって, その名詞句の指示対象を同定することが困難になる場合がある. それを表現 " N_1 の N_2 " に言い換えることで, 名詞句の意味を限定し, かつ, 字数を削減することが可能である. 言い換えは, 動詞を削除することによって行う. 表現 " N_1 の N_2 " では, 語 N_1 と N_2 の意味関係を示す述語が省略されている場合がある. この省略されうる述語を, 削除可能な動詞として 2 種類の方法により定義した. 一方では, 表現 " N_1 の N_2 " の意味構造に対応する動詞を, シソーラスを用いて選択した. また, 他方では, ある語から連想される動詞を定義した. ただし, コーパスから, 名詞とそれが係る動詞との対を抽出し, 共起頻度の高いものを, 名詞から動詞を連想可能であると考えた. これらの削除可能な動詞を用いた言い換えを評価したところ, 再現率 63.8%, 適合率 61.4%との結果を得た. さらに, 言い換え可能表現の絞り込みを行うことによって適合率は 82.9%に改善することが可能であることを示す.

キーワード: 言い換え,表現 " N_1 の N_2 ",修飾部削除,要約

Paraphrasing a Japanese verbal noun phrase into an expression " N_1 no N_2 "

Akira Kataoka † , Shigeru Masuyama †† and Kazuhide Yamamoto †††

The purpose of this paper is to propose a method of paraphrasing a Japanese verbal noun phrase into a noun phrase in the form of " N_1 no N_2 ". The semantic structure of " N_1 no N_2 " can be recognized by supplementing some abbreviated predicate. We define "deletable verbs" as these abbreviated predicates in two ways. 1. Choose verbs equivalent to the semantic relations of " N_1 no N_2 " using a thesaurus. 2. Choose verbs associated with nouns. If a verb frequently cooccurs with a noun in newspaper articles, it is concluded that the verb is associated with the noun. By defining "deletable verbs" and utilizing a variety of the semantic structure of " N_1 no N_2 ", this paraphrasing is accomplished. The subjective evaluation of our paraphrasing method shows that the precision is 63.8% and the recall is 61.4%. It is also shown that restriction on targets can increase the precision by 82.9%.

KeyWords: paraphrasing, expression "N₁ no N₂," deletion of modifiers, summarization

[†] 豊橋技術科学大学 知識情報工学系, Information Engineering, Toyohashi University of Technology(現在, 日本電信電話 (株) コミュニケーション科学基礎研究所, NTT Communication Science Laboratories)

^{††} 豊橋技術科学大学 知識情報工学系, Department of Knowledge-based Information Engineering, Toyohashi University of Technology

^{† † †} ATR 音声言語通信研究所,ATR Spoken Language Translation Research Laboratories

1 はじめに

本論文では、表現 " N_1 の N_2 " が多様な意味構造を持つことを利用して、動詞を含む連体修飾節を表現 " N_1 の N_2 " に言い換える手法を提案する.

自然言語では、一つの事象を表すために多様な表現を用いることが可能であり、人間は、ある表現を、同じ意味を持つ別の表現に言い換えることが、しばしばある。言い換えは、自然言語を巧みに操るために不可欠な処理であり(佐藤 1999)、それを機械によって実現することは有用であると考えられる。例えば、文書要約において、意味を変えずに字数を削減するためや、文章の推敲を支援するシステムにおいて、同一の表現が繰り返し出現するのを避けるために必要な技術である。また、ある事象が様々な表現で表されているとき、それらの指示対象が同一であると判定するためにも必要である。

8節で述べるように、近年、言い換え処理の重要性はかなり認識されてきたと考えられるが、適切な問題の設定を行うことが比較的困難なため、言い換え処理の研究はそれほど進んでいない。佐藤(佐藤 1999)は、「構文的予測の分析」から「構文的予測を分析する」への言い換えのように、動詞を含む名詞句を述語の形式に言い換える問題を設定している。また近藤ら(近藤、佐藤、奥村 1999)は、「桜が開花する」から「桜が咲く」への言い換えのように、サ変動詞を和語動詞に言い換える問題設定をしている。この他、「~を発表しました.」から「~を発表.」のような文末表現の言い換えや、「総理大臣」から「首相」のような省略形への言い換えなどを、言い換えテーブルを用意することによって実現している研究もある(若尾、江原、白井 1997; 山崎、三上、増山、中川 1998).

これに対し我々は、名詞とそれに係る修飾語、すなわち連体修飾表現を異形式の連体修飾表現に言い換えるという問題設定を提案する。2節に述べるように、我々は連体修飾表現を言語処理の観点から3分類し、これらの相互の変換処理を計算機上で実現することを研究の最終目標として設定し、このうち本論文において動詞型から名詞型へ変換する手法を議論する。連体修飾表現を対象にした本論文のような問題設定は従来見られないが、表現が短縮される場合は要約などに、また逆に言い換えの結果長い表現になる場合は機械翻訳などの処理に必要な処理であると考える。

本問題においても、従来研究と同様言い換えテーブルを用意することで言い換え処理を実現する.しかし本論文では、その言い換えテーブルを如何にして作成するかについて具体的に述べる.連体修飾表現の言い換え可能な表現は非常に多く存在することが容易に想像でき、これらをすべて手作業で作成することは現時点においては困難である.このため、現実的な作業コストをかけることで言い換えテーブルを作成する手法を示す。本提案処理の一部にはヒューリスティックスが含まれているが、これらについても一部を提示するにとどめず、具体例をすべて開示する.

本論文で言い換えの対象とする表現 " N_1 の N_2 " は, 2 つの語 N_1 , N_2 が連体助詞 'の' によって結ばれた表現である. 表現 " N_1 の N_2 " は, 多様な意味構造を持ち, さまざまな表現をそれに言

い換えることが可能である。また、動詞を含む連体修飾節は、各文を短縮する要約手法(三上、増山、中川 1999;山本、増山、内藤 1995)において削除対象とされている。しかし、連体修飾節すべてを削除することにより、その名詞句の指す対象を読み手が同定できなくなる場合がある。このとき、それを " N_1 の N_2 " という表現に言い換えることができれば、名詞句の指示対象を限定し、かつ、字数を削減することが可能となる。表現 " N_1 の N_2 " は多様な意味を持ちうるため、たとえ適切な言い換えがされたとしても、曖昧性が増す場合がある。しかしながら、言い換えが適切であれば、読み手は文脈や知識などを用いて理解が可能であると考えられる。

以下, 2 節で, 連体修飾表現を分類し, 本論文で対象とする言い換えについて述べる. 3 節から 5節で本手法について述べ, 6 節では主観的に本手法を評価する. 7 節では, 評価実験の際に明らかになった問題点などを考察する. また 8 節では, 本論文の関連研究について論じる.

2 連体修飾表現の言い換え

動詞型連体修飾 $N_1(m'/e')$ V-する N_2

名詞型連体修飾 N_1 の N_2

 $N_1 N_2$

形容詞型連体修飾 N_1 (が/を/…) A-い N_2

 N_1 ($m/e/\cdots$) A-a N_1

 $\overline{N_1,N_2}$:名詞, V-する:動詞, A-い, A-な:形容詞

図1 連体修飾表現の分類

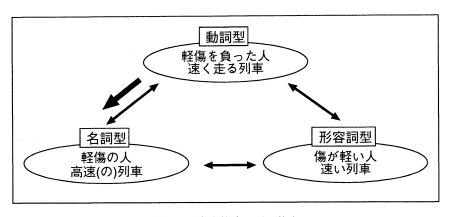


図2 型を変換する言い換え

本節では,連体修飾表現の言い換えという問題設定を,我々が如何にして行うかについて述べ,このうち本論文で対象とする問題の定義を行う.

まず自然言語処理の観点から、連体修飾表現の分類として、図 1 のように、動詞型、名詞型、形容詞型の 3 種類の型を定義する。名詞を修飾するという同じ役割に対して、さまざまな表現が可能である。例えば図 2 のように、異なる型においても、ほぼ同一の意味を持つ表現が存在する。人間ならば、これらの表現が同一の意味を持つと理解した上で、互いに言い換えることが可能である。そこで本研究では、これを計算機で行うことを目指す。すなわち、ある型の連体修飾表現を、如何にして他の型の連体修飾表現に変換するか、という型変換処理の形で問題を設定する。

機械によって、それぞれの表現の意味を理解し、言い換え可能であるか判定することは現状の技術では困難であり、何らかの表層表現を手がかりにした近似的な手法を考案する必要がある。よって我々は、それぞれの型変換における表層的な特徴を利用して言い換えを実現する.

図 2 のうち本論文で対象とするのは、動詞型連体修飾表現を名詞型連体修飾表現 " N_1 の N_2 " へ言い換えるものである。他の型の表現を " N_1 の N_2 " に言い換えることで、要約において、文の冗長さを減少させることや、文章推敲を支援するシステムにおいて、文章中に同一の表現が続くことを避けることなどが可能である。

例 1

A 国で起きたクーデター事件 → A 国のクーデター事件

動詞型連体修飾表現は、各文を短縮する要約手法 (三上他 1999; 山本他 1995) において削除対象となっている。しかし上の例 1 のように、連体修飾表現を削除すると、読み手が、その名詞句の指す対象を同定することが困難になる場合も存在する。こういった連体修飾表現を表現 " N_1 の N_2 " に言い換えることによって、可能な限り字数を削減することができ、かつ、その指示対象の同定を容易にすることができる。

本論文で提案する手法によって、" N_1N_2 "への言い換えも、同様に可能であると考える.

例 2

外国で製作された映画 → 外国の映画

→ 外国映画

しかし、この言い換えを行う場合、以下の点を判定する必要があり、本論文では、この言い換えは 扱わない。

• N_1, N_2 を連結し、複合名詞 " N_1N_2 " として扱うことが可能か.

例 3

太郎が持つ考え → 太郎の考え

→ (?) 太郎考え

• " N_1 の N_2 " が指す対象と、" N_1N_2 " のそれとが同一であるか.

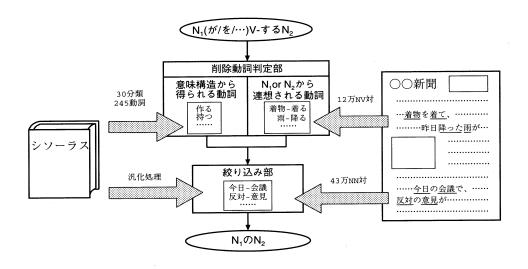


図 3 本手法の概念図

例 4

日本にある大学 → 日本の大学

→ (?) 日本大学

なお,本論文でいう動詞には、"サ変名詞 + する"を含み、"する"、"なる"、"である" など 1 を含まない。

3 本手法の構成

本手法は、以下に示す部分から成る. また、本手法の概念図を図3に示す.

削除動詞判定部 動詞型連体修飾表現に含まれる動詞が, 2 種類の方法で定義した削除可能な 動詞であるか否かを判定する.

言い換え表現絞り込み部 コーパスに出現しない表現 " N_1 の N_2 " に言い換えることがないよう言い換えに制限を加える.

表現 " N_1 の N_2 " の中には、語 N_1 と N_2 を結ぶ述語が省略されている、連体修飾節の短縮形と考えられるものが存在する (平井 北橋 1986; 黒橋 酒井 1999; 島津, 内藤, 野村 1985). この省略されうる述語として、本論文では図 3 に示すように 2 種類の "削除可能な動詞" を定義する (Kataoka, Yamamoto, and Masuyama 1999).

¹ その他は, "よる", "できる", "関する", "対する", "いう", "つく", "伴う".

表現 " N_1 の N_2 " の意味解析に関する既存の研究では,語 N_1 , N_2 の意味関係を幾つかのクラスに分類することを試みている (平井・北橋 1986; 国立国語研究所 1951). これらの意味構造に対応する動詞は,削除可能であると考えられる.一方,表現 " N_1 の N_2 " には, N_1 , N_2 間の意味関係を示す動詞を,語 N_1 , または, N_2 から連想できる場合がある.この場合,連想される動詞は,その語と共起したときのみ 表現 " N_1 の N_2 " の意味構造に対応するため,常に既存の分類と対応するとは限らない.どの動詞が連想されるかは,日常的な語の使われ方によって決まるため,その情報は,コーパスから得ることが適切である.よって,以下の 2 種類の方法により,削除可能な動詞を定義する.

- \bullet シソーラスを用いて、表現 " N_1 の N_2 " の意味構造に対応する動詞を選択する.
- コーパスから, 名詞と動詞の対を抽出し, 共起頻度の高い対の名詞から動詞が連想可能であると判定する.

これらの削除可能な動詞を用いることで、動詞を含む連体修飾節が表現 " N_1 の N_2 " に言い換えられることを示す。

一般に言語現象は複雑であり、問題解決のための規則を人間が記述する規則利用型 (rule-based) 処理において、すべての現象をとらえられる規則を記述するのは困難である. 一方、用例利用型 (example-based) 処理では、コーパスに類似した用例が出現しない場合、問題に対処することができない. これらの理由から、本論文では、2種類の方法によって削除可能な動詞を定義する.

4 削除可能な動詞

名詞型連体修飾 " N_1 の N_2 " の意味解析に関する研究は、従来から多く行われている (平井・北橋 1986; 国立国語研究所 1951; 黒橋・酒井 1999; 島津他 1985; 冨浦, 中村, 日高 1995). 平井ら (平井・北橋 1986) は、表現 " N_1 の N_2 " に、 N_1 と N_2 を結ぶ述語が省略されているものが存在するとしている。また、この種の " N_1 の N_2 " の意味を理解するためには、読み手が、省略された述語を推定できなければならないことから、それらの述語は非常に基本的な関係を示すものであるとしている。また、英語の複合名詞句において、2つの名詞間の意味関係を 9種の深層レベルの述語 RDP(Recoverably Deletable Predicate) 2 によってとらえる研究もある (Levi 1978).

これらの研究でも示されているように、" N_1 の N_2 " という表現や、複合名詞では、述語が省略されているものが存在する。逆に言うと、この省略されうる述語を含む連体修飾 " N_1 (が/ e/\cdots)V-する N_2 " は、表現 " N_1 の N_2 " に言い換えることが可能である。

以下の例 5 では、動詞 "発表する" を削除して、" N_1 の N_2 " と言い換えることができるが、動詞 "批判する"を削除して言い換えることはできない。これは、語 N_1 と N_2 が "発表する" によって結ばれた場合、その意味関係は " N_1 の N_2 " の意味構造に対応するが、"批判する" によって結 $\overline{2}$ 具体的には、CAUSE、HAVE、MAKE、USE、BE,IN、FOR、FROM、ABOUT の $\overline{9}$ 種.

ばれた場合、それに対応しないからだと考えられる.

例 5

首相が発表した法案 → 首相の法案 首相が批判した法案 → (?) 首相の法案

本論文では、この "発表する" のような省略されうる動詞を "削除可能な動詞" と呼び、2 種類の方法により定義する.

4.1 " N_1 の N_2 "の意味構造から得られる動詞

国立国語研究所は、" N_1 の N_2 "の意味構造を人手により分類している (国立国語研究所 1951). その中で、述語が省略されていると考えられる分類を抜き出し 3 、以下に示す.

• 所有主.

(例: 太郎のボール)

• 執筆者, 発信者, 主催者, 主演者など, 後ろの体言の作成行為をなした者.

(例:漱石の小説,首相の談話)

所属の団体.

(例:A社の役員)

存在の場所・位置.

(例:奈良の東大寺)

これらの意味関係を示す動詞を,削除可能な動詞であると考える.また,削除可能な動詞は, 平井らが述べているように「非常に基本的な関係を示す述語」でなければならないため,新聞記 事に出現する頻度が上位の動詞だけを含める.

よって、以下の条件1を満たす動詞を削除可能な動詞であると定義する.

条件1:

- シソーラス (大野 浜西 1981) において, 上記の意味関係を示すと考えられる分類に含まれる. かつ,
- コーパスに出現した動詞のうち、出現頻度が上位である。

上記の意味関係に対応するシソーラスの分類としては、"所有"、"生成"、"開始"、"表現"、 "実行"、"生産"など、末端の分類で30分類を選択した。また、コーパスとして日本経済新聞 1993年の全記事を使用し、それらに出現した動詞(約2万語)を観察した結果、上位10%に当た る2000位以上の動詞を出現頻度が上位であると判断した。

その結果, 削除可能な動詞を, "発表 (する)", "始める", "まとめる", "開く", "実施", "決める", "開始", "建設", "行う" など, 245 個登録した. 付録 A に, 選択した 30 分類, および, 削除可能な動詞の一部を示す.

³ なお、これらの分類は、必ずしも並列ではない。

なお, これらの動詞の中には, 削除可能な動詞として不適切な動詞 "偽造 (分類:製造)", "冷蔵 (分類:保有)" なども含まれているが, 客観性を保つため, それらを人手で除去することは行わなかった.

4.2 語から連想される動詞

前節では、削除可能な動詞をシソーラスを用いて定義した。しかし、これら以外の動詞であっても、文脈によって削除可能となる場合がある。以下の例 6 では、動詞 "着る" や "降る" を削除して " N_1 の N_2 " と言い換えることができる。

例 6

着物を着た女性 → 着物の女性

雨が降った日 → 雨の日

これは、名詞 "着物"、"雨" から、それぞれの動詞を連想できるためと考えられる。" N_1 の N_2 " の意味解析に関する研究 (平井・北橋 1986; 黒橋・酒井 1999; 島津他 1985; 田中、冨浦、日高 1998) においても、語 N_1 、または、 N_2 から連想される動詞を補完することで、その意味関係がとらえられる場合があるとしている。

これらの動詞は、前述の定義では削除可能な動詞として定義されない.これらの動詞を削除 可能であると判定するためには、ある名詞から連想される動詞を判定する必要がある.そこで、 新聞記事において、ある名詞と、それが係る動詞との対を抽出する.ある名詞が与えられたとき、 抽出した対の中で、その名詞と共起頻度の高い動詞を連想される動詞であると判定する.

NV 対の抽出

以下の手順により、新聞記事から名詞と動詞との対を抽出する. この抽出される対を、"NV 対 "と定義し 4 、"(n,v)"と表記する.

- (1) 記事に対して形態素解析、および構文解析を行う. 形態素解析器は JUMAN⁵を用いた. また、構文解析器は、Perl 言語を用いて独自に実装し、基本的に、すべての助詞は最も近い後方の用言に係ると判定した. 解析結果の人手による修正は行わない.
- (2) 解析結果から, 以下を NV 対 として抽出する.
 - 動詞 (v) と, その格要素の主辞 (n)
 - 修飾表現内の動詞 (v) と, その被修飾部の主辞 (n)

ただし、本論文では、名詞、または、接尾辞が連続している部分のうち最も後方の形態素を、主辞と定義する。 また、n の品詞 (JUMAN の解析結果) が数詞、人名、地名、組織名のいずれかであるならば、それぞれの品詞名をn として抽出する.

⁴ 田中ら (田中他 1998) は、一部の " N_1 の N_2 " の意味推定の際に、本論文と同様、コーパスにおける名詞と動詞の共起関係を用いている。

⁵ http://pine.kuee.kyoto-u.ac.jp/nl-resource/

例えば、"東京で開かれた国際会議に出席する." という文からは、〈地名、開く〉、〈会議、開く〉、〈会議、出席〉という NV 対が抽出される.

NV 対 を抽出する際に, v に対して n がとる格を考慮することが考えられる. ところが, 連体修飾表現 "V-する N_2 " においては, 被修飾語 N_2 がとる格を表層表現から得ることができない. また, 文型によって表層格が変化しても同一の NV 対 として抽出することが望ましい. これらの理由から, 表層的な情報のみを扱う本論文では, NV 対 において, 格の情報は扱わない.

また,前述の抽出法では,連体修飾表現が「外の関係」⁶である場合,動詞の格要素ではない名詞が抽出される.しかし一般に,ある連体修飾表現が「内の関係」であるか「外の関係」であるかを機械的に判定することは困難であるため,本論文ではその判定は行わない.

日本経済新聞 1993 年の全記事 (約 15 万記事) に対して抽出を行った結果, 約 470 万の NV 対が抽出された (表 1).

表 1 NV 対 の抽出結果

抽出対象 約15万記事 NV 対の延べ数 約470万 NV 対の異なり数 約140万 Nの異なり数 約62,000 Vの異なり数 約15,000

NV 対による削除動詞の判定

抽出された NV 対 を用いて, 連想される動詞の判定を行う.

まず、名詞 n と動詞 v の共起率 $CR_n(v)$ を次式によって定義する.

$$CR_n(v) = \frac{F(n, v)}{\sum_{for\ all\ i} F(n, v_i)}$$

 $F(n,v):\langle n,v \rangle$ の出現頻度

ある名詞に対して最も高い共起率を持つ動詞を、連想される動詞として定義する。 つまり、 $\langle n,v \rangle$ が以下の条件 2 を満たすとき 7 、名詞 n から動詞 v が連想されると判定する.

条件 2:
$$\forall i, CR_n(v) \geq CR_n(v_i)$$

前節で抽出された NV 対 に対して判定を行った結果, 異なり数で約 12 万対が条件 2 を満たし, 名詞から動詞が連想されると判定された. 表 2 に, 条件 2 を満たす NV 対 について延べ数など

⁶ 被修飾語が修飾部の用言の格要素とならない (寺村 1975 1978). 格要素となる場合を「内の関係」と呼ぶ.

⁷条件2は、 $CR_n(v)$ を用いず、F(n,v)によっても同等の定義が可能である.

を示し、付録 B にその例を示す。表 2 において、N の異なり数と、NV 対 の異なり数とが一致していない。これは、ある名詞に対して、条件を満たす NV 対 が複数存在する場合、それらの動詞すべてを、その名詞から連想される動詞として判定しているからである。

表 2 条件 2 を満たす NV 対

約 67 万
(全 NV 対 の 14.2 %)
約12万
約 62,000
約 7,000

ある動詞が、4.1節、および、本節の 2 つの方法で重複して削除可能と判定される場合もある。 条件 2 を満たす 12 万対のうち、シソーラスを用いた定義によっても削除可能と判定されるもの は約 2 万対であった。重複して判定される動詞の例を表 3 に示す。表中の NV 対 の動詞は、シ ソーラスを用いても削除可能であると判定される。

表 3 重複して削除可能と判定される動詞

NV 対	頻度	$CR_n(v)$
〈社債,発行〉	230	44.8 (%)
〈結論,出す〉	496	40.6
〈会議,開く〉	1995	24.8
〈教書,発表〉	11	19.6
【 (伸び, 示す)	468	13.9

NV 対の抽出、および、それによる削除動詞の判定において、4.2 節に述べた理由から、名詞と動詞の格関係を考慮していない。そのため、以下の状況が生じうる。まず、 $CR_n(v,c)$ を、名詞 n と動詞 v が格関係 c によって係る割合e とする。ある e0、が、条件 e2 を満たすとしても、e0、のよい。を最大にする格関係 e2 において、e1、e2 において、e3 において、e4、e5 である e7 が存在する可能性がある。格関係が異なる e7、e8 は、異なる意味関係で共起していると考えることもでき、この状況では、e7 から連想される動詞として e7 が適切であるとは限らない。

しかし、 $\langle n,v\rangle$ が条件 2 を満たす際には、ある特定の格関係が $CR_n(v)$ に大きく寄与していると考えられる。 すなわち、 $CR_n(v,c_i)$ を最大にする格関係 c において、 $CR_n(v,c) \simeq CR_n(v)$ となることが多く、前述の状況は生じにくい。例えば、 \langle 犯人、逮捕 \rangle という例では、その格関係は ヲ格 (対象) のみと考えるのが自然である。本論文において、名詞から動詞が連想されると判定された NV 対 を観察したところ、名詞と動詞の格関係は一定の場合が多かったことから、前述

⁸ 分母は, $CR_n(v)$ の定義と同様, n の出現頻度.

の状況となる NV 対 は少ないと考えられる.

一方,複数の格関係が $CR_n(v)$ に寄与しうる例として、〈本、読む〉が挙げられる。ところが、深層格が異なる"本を読む"、"本で読む"という表現において、同一のNV 対が抽出され、"本"から"読む"が連想されると判定されても問題はない。格関係が異なっているとしても、係り受け関係を持って共起していることから、連想される動詞としての意味的な関係は、ある程度妥当である場合が多いと考えられる。

以上の議論から、本論文において NV 対 に格の情報を含めなかったことが、精度に与える影響は小さいと予想する. 4.2 節 で述べた理由によって NV 対 のデータ量を確保するという観点から、格の情報を考慮しないことが現実的には有利な選択であろう.

5 言い換え可能表現の絞り込み

連体修飾表現の動詞が慣用句の一部である場合など、たとえ動詞が削除可能な動詞であっても、" N_1 の N_2 "と言い換えると不自然となることがある。以下の例 7 では、慣用句 "力を入れる"の動詞を削除して言い換えると意味が分からない表現となる。

例 7

力を入れる交渉 → (?) 力の交渉

また, V が同じ語であっても, N_1 , N_2 の格が異なれば, 言い換えが不自然になる例がある.

例 8

裁判長が出した勧告 → 裁判長の勧告

勧告を出した裁判長 → (?) 勧告の裁判長

このような不自然な言い換えを避けるため、動詞型連体修飾 " N_1 (が/e/e/e/ \cdots) V-する N_2 " の言い換えにおいて、語 N_1 、 N_2 がコーパス中に " N_1 の N_2 " の形で出現している場合にのみ、言い換えを行う。例 7,8 では、言い換え後の表現である "力の交渉"、"勧告の裁判長" は、コーパスに出現しないと考えられることから、不自然な言い換えを回避できる。

5.1 NN対の抽出

コーパス中に " N_1 の N_2 " の形で出現する名詞句に含まれる語 $N_1,\,N_2$ の対を NN 対 として定義し, $[n_1,\,n_2]$ と表記する.

新聞記事から,以下の手順により NN 対 を抽出する.

- (1) 記事に対して JUMAN による形態素解析を行う. 解析結果の人手による修正は行わない.
- (2) 接続助詞 'の' による修飾表現のうち, 修飾部の主辞 (n_1) と, 被修飾部の主辞 (n_2) を抽出する. ただし, n_1 , n_2 の品詞 (JUMAN による解析結果) が人名, 組織名, 地名, 数詞

のいずれかの場合は、それぞれ品詞名を n_1 、あるいは、 n_2 として抽出する.

例えば、"形態素解析の実行結果"という表現からは、[解析、結果]という NN 対 が得られる. また、" N_1 の N_2 の N_3 "という形の表現において、語 N_1 が N_2 に係るか、あるいは、 N_3 に係るかを、表層的な情報から判定することは困難である. NN 対は、言い換えの際の誤りを排除するという目的から、正しいもののみが収集されていることが望ましい。よって、" N_1 の N_2 の N_3 "という形の表現からは、NN 対として $[N_2, N_3]$ のみを抽出する.

日本経済新聞 1993 年の記事から抽出を行った結果, 延べ数で約 105 万, 異なり数で約 43 万の NN 対 が抽出された.

5.2 NN対の汎化処理

日本語では、名詞のうち多くのものを接続助詞 'の' によって結合することができ、その結果、多くの表現を生成することが可能である.よって、コーパスから抽出した NN 対 データのスパース性が問題となる.これには、コーパス量を増やすことで対応することも考えられるが、本論文では、NN 対 に対してシソーラスを用いた汎化を行う.NN 対 の各語を、シソーラス (大野・浜西 1981) 中の末端の分類に置き換えた.その際、複数の意味カテゴリに分類されている単語は、各分類ごとに汎化した NN 対 を作成し、また、シソーラスに記載されていない単語に対しては、汎化を行わない.

6 評価

6.1 評価方法

日本経済新聞 1994 年の記事から, 動詞型連体修飾 " N_1 (が $/e/(e/\cdots)$ V-する N_2 " を人手で抽出し, 本手法の有効性を検討する. 本実験では, 動詞 V が格要素を一つ取っている表現のみを対象とした. これは, 以下の理由による. 例えば, " N_1 が N_2 を V-する N_3 " という表現を言い換える際に, " N_1 の N_3 ", " N_2 の N_3 ", " N_1 の N_2 の N_3 " のいずれの表現に言い換えるか, 文脈に応じて適切な表現を選択する必要がある. 本論文は, 言い換えを行う際に削除できる動詞を判定する手法を提案するものであり, いずれの格要素を残すべきかの選択は対象外とする.

記事より,動詞型連体修飾表現を無作為に 500 個抽出した. これらの表現に対して,人間,および,本手法によって言い換えを行い,再現率,適合率で評価する.

再現率: $R = C/H \times 100(\%)$, 適合率: $P = C/M \times 100(\%)$

ここで, H は, 筆者が主観によって, " N_1 の N_2 " に言い換えられるかどうかを判定し, 言い換え可能と判定された表現の数を示す。また, M は, 本手法によって言い換え可能と判定された表現の数を示し、C は, 人間と本手法とで共に言い換え可能と判定された表現の数を示す。

	H	M	C	R~(%)	P~(%)
制限無し	152	158	97	63.8	61.4
制限 1	152	(93,86)	(61,53) 29	19.1	82.9
制限 2	152	(27,16) 83	(23,13) 64	42.1	77.1
		(57,40)	(44,32)		

表 4 評価結果

表 5 正しく判定された例

言い換え前	言い換え後	動詞の分類	NV 対	$CR_n(v)$
高シェアを持つ会社	高シェアの会社	所有		
組合で作る連合会	組合の連合会	生成		
一日に開く会議	一日の会議	挙行,開始	〈会議,開く〉	24.8
大賞を受賞した A さん	大賞の A さん		〈大賞,受賞〉	23.5
賛成に回る議員	賛成の議員	_	〈 賛成, 回る 〉	22.7
低迷が続く業績	低迷の業績		〈低迷,続く〉	18.3

NN 対による絞り込みでは、以下の3種類の制限を用いて実験を行った.

- (制限無し)制限を設けない
- (制限 1) [n₁, n₂] の頻度が 1 以上ならば言い換える
- (制限 2) 汎化した [n₁, n₂] の頻度が 1 以上ならば言い換える

6.2 評価結果

表 4 に、評価結果を示す。 表中の M, C 列に示す括弧で括られた 2 つの数 (a,b) は、それぞれ、

- a: シソーラスを用いて定義された動詞によって言い換えられた表現の数
- b: 連想可能な動詞と判定された動詞によって言い換えられた表現の数を表す.

また, 削除可能な動詞が正しく判定され, 言い換えられた動詞型連体修飾の例 (制限無し) を表 5 に示す. シソーラスを用いた定義により判定されたものには動詞の意味分類を示し, 連想可能と判定されたものには NV 対と $CR_n(v)$ を示す.

7 考察

本論文では、動詞の表層的な情報のみに基づく判定によって、" N_1 の N_2 " への言い換えを行ったが、NN 対 による制限を加えない場合、再現率 63.8 %、適合率 61.4 % とおおむね良好な結果が得られた。

本手法では、2種類の方法により削除可能な動詞を定義した.シソーラスによる定義のみで、

あるいは, 連想される動詞のみで言い換えを行うと仮定すれば, それぞれ再現率が 30-40% 程度となることから, 2 種類の方法を併用して定義したことが有効であったといえる.

まず, NN 対 による制限無しの場合に, 再現率, 適合率を低下させた原因について考察する. 再現率を下げた原因には, 以下のことが挙げられる.

- "手掛ける", "抱える" など, 新聞記事においては"実施", "所有" などの意味を示しうるが、4.1節で選択したシソーラスの分類には含まれていない動詞があった.
- *NV* 対 は比較的高頻度で出現するが, その名詞に対して最も共起頻度が高い動詞ではなかったため, 連想される動詞として判定されなかった.

再現率を上昇させるために、シソーラスを用いた削除可能な動詞の定義において、選択する意味分類を、対象コーパスに適応させることが考えられる。しかし、コーパスにおける動詞の使用状況を調査する必要があるなど、その実現は容易ではない。もちろん、実験によって発見された動詞を、削除可能な動詞として新たに加えることは可能である。また、NV 対 による削除動詞の判定において、高い $CR_n(v)$ を持つ NV 対 を採用することによっても再現率の上昇が期待できる。しかし、その閾値は実験により求める他になく、決定は困難である。

適合率を低下させた原因には、以下のことが挙げられる. これらは、*NN* 対 を用いた絞り込みによっても排除することができない.

- 本実験では、新聞記事から NV 対 を抽出した。そのため、〈前年、上回る〉、〈経費、削減〉といった NV 対 の出現頻度が高くなった。これらは、新聞記事において頻出するが、その名詞から動詞が連想可能とは言えない。 NN 対 を用いた制限を行っても、例えば、"前年の成績" という表現がコーパスに出現していれば、"前年を上回る成績" をそれに言い換えてしまう。この問題に対しては、新聞記事に限定せず、多様なコーパスから NV 対 を抽出することで避けられると考えている。
- 〈質問,答える〉、〈費用,かかる〉なども出現頻度が高く、直観的に名詞から動詞が連想可能であると言える.しかし、以下のような動詞型連体修飾として出現した場合、動詞を削除すると意味が変化する.

例 9

質問に答えた結果 → (?) 質問の結果 費用がかかる調査 → (?) 費用の調査

ところが、以下のような文脈を考えることで、同様の言い換えは許容されると考えられる.

例 10

A は、B の質問に答えた結果、 \cdots \rightarrow A は、B の質問の結果、 \cdots

莫大な費用がかかる調査 \rightarrow 莫大な費用の調査 よって、これらの例は、本研究における評価では失敗としたが、NV 対 から得られる連想

可能な動詞に対する反例であるとは考えていない。実際には、" N_1 の N_2 " は単独で出現するのではなく、必ず文章中の他の語と共起して出現するため、文脈を考慮した判定、評価が必要である。しかしながら、考慮に入れるべき文脈の範囲を決定することは容易でなく、また、現在の技術では、正確な文脈解析を期待できない。したがって本論文では、修飾表現内で観測可能な現象のみを対象とした。

なお,シソーラスを用いて定義された動詞が原因で,不適切な言い換えを行い, NN 対 による制限によっても排除できなかった例も存在する.上述したように,実験によって発見された,これらの動詞を除くことは可能である.

次に、NN 対 による制限を加えた場合について考察する。制限を加えることで、適合率が上昇し、言い換えの誤りを除くという目的を達成することはできた。しかし、その一方で再現率が減少する。制限によって再現率が大幅に減少するのは、NN 対 データのスパース性が影響しているためとも考えられる。しかし、コーパス量を 2 倍 (日本経済新聞 2 年分) として NN 対を抽出しても再現率は数%程度しか上昇しなかった。したがって、NN 対を汎化する際に意味レベルをどのように設定するかの影響が強いと考えられる。最適な汎化レベルを求めることは、シソーラスの編集方針に依存するため容易ではない。また新聞記事では、

- " N_1 の N_2 " に言い換え可能な表現は、初めからそれで表現される、
- " N_1 の N_2 " によって表現すると曖昧さが残るものは、動詞型連体修飾で表現されると考えられる。そのため、動詞型連体修飾表現に出現する N_1 , N_2 と、" N_1 の N_2 " との間に重複が少なく、適切な NN 対 が収集されなかった可能性がある。

また, NN 対 を汎化する際, 複数の意味カテゴリに分類されている名詞は, 各分類ごとに汎化を行った. この汎化処理では, 名詞が, その意味では使用されていないカテゴリへ誤って汎化される恐れがある. しかし, 誤った汎化を行ったとしても, それに対応する表現が絞り込みの対象とならない限り影響はない. 実験では, 汎化処理を行った絞り込み (制限 2) における適合率の減少は 5%程度と高くはないことから, 誤った汎化の影響を受ける " N_1 の N_2 " が絞り込みの対象となる確率は低いと考えられる. また, 汎化を行った結果, 再現率が 20%程度改善されている. NN 対 データのスパース性に対処するという目的で汎化を行っており, また, 適合率の減少は再現率の上昇と比較して微小であることから考えて, 誤った汎化の回避は, 優先して取り組むべき課題ではないと考えている.

8 関連研究

まず、言い換えに関する既存の研究について論じる. 1節で挙げた研究の他では、加藤ら (加藤 浦谷 1999) は、原文とその要約文との対応がとれたコーパスを用いて、言い換えが行われている部分を照合により特定し、それを言い換えの知識として自動的に得る手法を提案している. また、Hovy ら (Hovy and Lin 1997)、近藤ら (近藤 奥村 1996) は、シソーラスを用いて、意味が類

似した複数の語句を、より抽象的な一つの語句に言い換える手法を提案している。また、近藤ら (近藤、佐藤、奥村 2000) は、「犬が彼に噛み付く」から「彼が犬に噛み付かれる」のような、単 文中の非が格要素をが格化する言い換えを実現するための規則を提案している。連体修飾節を 対象とした言い換えに関しては、これまで、ほとんど研究されていなかったが、野上ら (野上、藤田、乾 2000) によって、「1742 年に創立されたコスタは、スウェデーン最古の工場だ。」から「コスタは、スウェデーン最古の工場だ。1742 年に創立された。」への言い換えのように、連体修飾節を主文から切り離す言い換えが取り上げられている。

次に、本論文における削除可能な動詞、および、その定義に関連する研究について論じる. 田中ら (田中他 1998) は、" N_1 の N_2 " の意味関係を推定するために、「一般的な意味関係」 9 と「名詞固有の意味関係」を定義している。この中で、本論文における NV 対の定義、および、それによる連想される動詞の判定は、田中らが「名詞固有の意味関係」を得る際に行う処理とほぼ同一である。田中らの概念は、冨浦ら (冨浦他 1995) による意味推定において曖昧性が残る " N_1 の N_2 " を対象としている。一方、本論文では、言い換えを行う際の削除動詞を決定するという立場から、言い換え可能な " N_1 の N_2 " すべてを対象としている。そのため、「連想される動詞」によって言い換えられる " N_1 の N_2 " は、田中らの「名詞固有の意味関係」と、必ずしも一致しない。また田中らの概念では、意味推定という立場から、対象としている " N_1 の N_2 " は、2 つの概念のいずれかに分類される。一方、本論文における 2 つの概念は互いに排他的ではなく、4.2 節で議論したように、両者の概念によって定義される動詞も存在する。これらの相違は、田中らが意味推定、本論文においては言い換え、と異なる目的のために 2 つの概念を定義し、利用していることにあるといえる。

また、村田ら(村田 長尾 1998)、山本ら(山本、村田、長尾 1998)は、名詞や動詞の省略補完において、本論文と同様に、コーパスから取得した用例を利用している。ただし、村田らの手法(村田・長尾 1998)では、コーパスに対して形態素解析や構文解析をせず、単なる文字列として最長に一致する部分を用例と認定している。また山本らの手法(山本他 1998)では、名詞と動詞との係り受け関係に関する情報は格フレーム辞書から得ている。これらの手法も、田中らと同様、表層的には存在しない動詞を推定することを目的とする。よって、用例を利用している点では本論文と類似するが、目的は異なる。

9 おわりに

本論文では、動詞型連体修飾 " N_1 (が/ $e/e/\cdots$) V-する N_2 " e, 名詞型連体修飾 " N_1 の N_2 " に言い換える手法を提案した.

" N_1 の N_2 "の中には、動詞型連体修飾において動詞が省略された短縮形と考えることができるものがあり、その省略されうる動詞を削除可能な動詞として 2 種類の方法によって定義した.

^{9 「}一般的な意味関係」は、さらに、7つに分類される.

これらの削除可能な動詞を利用することで、動詞の表層的な情報のみを利用して、" N_1 の N_2 " への言い換えが実現可能であることを示した。また、コーパスに " N_1 の N_2 " の形で出現するもののみを言い換えることで、削除可能な動詞の判定の際の誤りを排除し、適合率を上げることが可能であることを示した。

今後の課題として, 文脈を考慮して削除可能な動詞を判定すること, 複数の格要素を持つ動詞型連体修飾表現を言い換えること, などが挙げられる.

铅憶

本研究で使用した「角川類語新辞典」を機械可読辞書の形で提供いただき、その使用許可をいただいた(株)角川書店に深謝する。また、本研究で言語データとして使用した日経新聞CD-ROM、1994年版の使用許可をいただいた(株)日本経済新聞社に深謝する。

付録

A シソーラスにより定義された動詞の例

分類	動詞
所有	共有, 持つ, 所有, 占める, 占領, 独占, 備える
保有	確保,保つ,保管,保有,冷蔵
生成	形成, 結ぶ, 結晶, 構成, 作り出す, 作る, 作成, 成り立つ, 成る, 生じ
	る, 生まれる, 生み出す, 生む, 組み立てる, 創作, 創造, 造る, 誕生,
	発生,編成
挙行	開く, 開催, 挙げる, 共催, 行う, 催す, 執行, 主催
建造	改築, 建つ, 建てる, 建設, 建造, 建築, 構える, 構築, 再建, 新築, 組
	み立てる, 増築, 築く
存在	既存, 共存, 潜在, 存在, 分布

その他の分類

従属, 発生, 開始, 進捗, 提示, 表現, 叙述, 描写, 書き, 執筆, 発言, 言明, 総括, 実行, 遂行, 設置, 設備, 生産, 製造, 架設, 決定, 施設, 発表, 発行

B 連想可能な動詞の例

NV 対	頻度	$CR_n(v)$
〈平行線,たどる〉	137	88.3 (%)
〈注目,集める〉	781	73.8
(けじめ,つける)	98	67.5
〈長期間,わたる〉	57	57.5
〈ボタン,押す〉	86	50.0
〈汗,流す〉	144	41.0
〈 役割, 果たす 〉	1155	39.5
〈損害,与える〉	131	38.1
〈たばこ,吸う〉	73	33.4
〈白紙,戻る〉	53	29.6
くうわさ,流れる〉	131	27.9
〈賞,受賞〉	265	25.5
〈被害,受ける〉	451	24.1
〈材,使う〉	169	18.9
〈小説,書く〉	53	16.0
〈赤字,転落〉	494	15.1
〈治療,受ける〉	88	14.9
〈メッセージ,送る〉	62	14.9
〈役,務める〉	178	12.9
〈抵抗,あう〉	52	11.8
〈画面,表示〉	93	11.0

参考文献

- 平井誠 北橋忠宏 (1986). "日本語文における「の」と連体修飾の分類と解析." 情報処理学会研 究報告 NL-58-1.
- Hovy, E. and Lin, C.-Y. (1997). "Automated text summarization in SUMMARIST." In *Proc.* of the ACL Workshop on Intelligent Scalable Text Summarization, pp. 18–24.
- Kataoka, A., Yamamoto, K., and Masuyama, S. (1999). "Summarization by Shortening Japanese Noun Modifiers into Expression "A no B"." In *Proceedings of NLPRS99*, pp. 409–414.
- 加藤直人 浦谷則好 (1999). "局所的要約知識の自動獲得手法." 自然言語処理, 6 (7), 73-92.

- 国立国語研究所 (1951). 現代語の助詞・助動詞 -用法と実例-. 秀英出版.
- 近藤恵子 奥村学 (1996). "言い替えを使用した要約の手法." 情報処理学会研究報告 NL-116-20, pp. 137-142.
- 近藤恵子, 佐藤理史, 奥村学 (1999). "「サ変名詞+する」から動詞相当句への言い換え."情報 処理学会論文誌, **40** (11), 4064-4074.
- 近藤恵子, 佐藤理史, 奥村学 (2000). "格変換による単文の言い換え." 情報処理学会研究報告 NL-135-16, pp. 119-126.
- 黒橋禎夫 酒井康行 (1999). "国語辞典を用いた名詞句「A の B」の意味解析." 情報処理学会研 究報告 NL-129-16, pp. 109-116.
- Levi, J. N. (1978). The Syntax and Semantics of Complex Nominals. Academic Press.
- 三上真, 増山繁, 中川聖一 (1999). "ニュース番組における字幕生成のための文内短縮による要約." 自然言語処理, $\mathbf{6}$ (6), 65-81.
- 村田真樹 長尾真 (1998). "日本語文章における表層表現と用例を用いた動詞の省略の補完." 自然言語処理, **5** (1), 119-133.
- 野上優, 藤田篤, 乾健太郎 (2000). "文分割による連体修飾節の言い換え." 言語処理学会 第 6 回 年次大会 発表論文集, pp. 215-218.
- 大野晋 浜西正人 (1981). 角川類語新辞典. 角川書店.
- 佐藤理史 (1999). "論文表題を言い換える." 情報処理学会論文誌, 40 (7), 2937-2945.
- 島津明, 内藤昭三, 野村浩郷 (1985). "日本語文意味構造の分類 -名詞句構造を中心に-." 情報処理学会研究報告 NL-47-4.
- 田中省作, 冨浦洋一, 日高達 (1998). "統計的手法を用いた名詞句「NP の NP」の意味関係の抽出法." 電子情報通信学会技術研究報告 NLC-98-4, pp. 23-29.
- 寺村秀夫 (1975-1978). "連体修飾のシンタクスと意味 (1)-(4)." 日本語・日本文化 vol.4-7. 大阪外国語大学研究留学生別科.
- 冨浦洋一, 中村貞吾, 日高達 (1995). "名詞句「NP の NP」の意味構造." 情報処理学会論文誌, 36 (6), 1441-1448.
- 若尾孝博, 江原暉将, 白井克彦 (1997). "テレビニュース番組の字幕に見られる要約の手法." 情報処理学会研究報告 NL-122-13, pp. 83-89.
- 山本専,村田真樹,長尾真 (1998). "用例による換喩の解析." 言語処理学会 第 4 回年次大会 発表論文集, pp. 606-609.
- 山本和英, 増山繁, 内藤昭三 (1995). "文章内構造を複合的に利用した論説文要約システム GREEN." 自然言語処理, **2** (1), 39-56.
- 山崎邦子, 三上真, 増山繁, 中川聖一 (1998). "聴覚障害者用字幕生成のための言い替えによる ニュース文要約." 言語処理学会 第 4 回年次大会 発表論文集, pp. 646-649.

略歴

- 片岡 明: 2000 年 豊橋技術科学大学大学院工学研究科知識情報工学専攻修士課程修了. 同年, 西日本電信電話 (株) 入社. 現在, 日本電信電話 (株) コミュニケーション科学基礎研究所勤務. 在学中は,自然言語処理,特にテキスト要約の研究に従事. 現在は,機械翻訳の研究に従事. E-mail: kataoka@cslab.kecl.ntt.co.ip
- 増山 繁: 1977年 京都大学工学部数理工学科卒業. 1982年 同大学院博士後期課程単位取得退学. 1983年 同修了(工学博士). 1982年 日本学術振興会奨励研究員. 1984年 京都大学工学部数理工学科助手. 1989年 豊橋技術科学大学知識情報工学系講師, 1990年 同助教授, 1997年 同教授. アルゴリズム工学, 特に, 並列グラフアルゴリズム等, 及び, 自然言語処理, 特に, テキスト自動要約等の研究に従事. 言語処理学会, 電子情報通信学会, 情報処理学会等会員. E-mail: masuyama@tutkie.tut.ac.jp
- 山本 和英: 1996 年豊橋技術科学大学大学院博士後期課程システム情報工学専攻修了. 博士 (工学). 1996 年~2000 年 ATR 音声翻訳通信研究所客員研究員, 2000 年~ATR 音声言語通信研究所客員研究員, 現在に至る. 1998 年中国科学院自動化研究所国外訪問学者. 要約処理, 機械翻訳, 韓国語及び中国語処理の研究に従事. 1995 年 NLPRS'95 Best Paper Awards. 言語処理学会, 情報処理学会, ACL 各会員. E-mail: yamamoto@slt.atr.co.jp

(1999 年 11 月 4 日 受付) (2000 年 3 月 17 日 再受付) (2000 年 6 月 30 日 採録)